New genes from old: asymmetric divergence of gene duplicates and the evolution of development

doi: info:doi/10.1098/rstb.2015.0480

PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Review

Cite this article: Holland PWH, Marlétaz F, Maeso I, Dunwell TL, Paps J. 2017 New genes from old: asymmetric divergence of gene duplicates and the evolution of development. *Phil. Trans. R. Soc. B* **372**: 20150480.
http://dx.doi.org/10.1098/rstb.2015.0480

Accepted: 22 June 2016

One contribution of 17 to a theme issue 'Evo-devo in the genomics era, and the origins of morphological diversity'.

Subject Areas:
developmental biology, genomics, evolution

Keywords:
tandem duplication, genome duplication, homeobox, Lepidoptera, Mollusca, Mammalia

Author for correspondence:
Peter W. H. Holland
e-mail: peter.holland@zoo.ox.ac.uk

New genes from old: asymmetric divergence of gene duplicates and the evolution of development

Peter W. H. Holland[1], Ferdinand Marlétaz[1,2], Ignacio Maeso[1,3], Thomas L. Dunwell[1] and Jordi Paps[1,4]

[1]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK
[2]Molecular Genetics Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan
[3]Centro Andaluz de Biología del Desarrollo, Consejo Superior de Investigaciones Científicas/Universidad Pablo de Olavide, 41013 Sevilla, Spain
[4]School of Biological Sciences, University of Essex, Colchester, Essex, UK

PWHH, 0000-0003-1533-9376

Gene duplications and gene losses have been frequent events in the evolution of animal genomes, with the balance between these two dynamic processes contributing to major differences in gene number between species. After gene duplication, it is common for both daughter genes to accumulate sequence change at approximately equal rates. In some cases, however, the accumulation of sequence change is highly uneven with one copy radically diverging from its paralogue. Such 'asymmetric evolution' seems commoner after tandem gene duplication than after whole-genome duplication, and can generate substantially novel genes. We describe examples of asymmetric evolution in duplicated homeobox genes of moths, molluscs and mammals, in each case generating new homeobox genes that were recruited to novel developmental roles. The prevalence of asymmetric divergence of gene duplicates has been underappreciated, in part, because the origin of highly divergent genes can be difficult to resolve using standard phylogenetic methods.

This article is part of the themed issue 'Evo-devo in the genomics era, and the origins of morphological diversity'.

1. Background

The central goal of evolutionary developmental biology is to understand how evolutionary modification of developmental processes leads to morphological or physiological differences between populations, species and higher taxa. Ultimately, it should be possible to trace these developmental differences to genetic mutations or possibly epigenetic changes that occurred in evolution. Fundamentally, there are two alternative approaches used in the field, which we call the 'classical evo-devo' approach and the 'reverse evo-devo' approach, by analogy to classical and reverse genetics. In the classical approach, trait differences of interest are identified between two species or populations and then experimental approaches are designed to track down the underlying genetic differences responsible. The reverse evo-devo approach is fundamentally different. Instead of starting with a trait of interest, one starts with differences in genes or genetic organization and then searches for what effect these differences could have on downstream phenotype. If the molecular differences are associated with genes thought to have roles in development (such as spatio-temporally regulated genes encoding transcription factors or signalling molecules), then it is a reasonable assumption that these differences will be associated with phenotypic differences in development. One would hope that the two approaches might ultimately converge, but the field is far from that

THE ROYAL SOCIETY PUBLISHING

state at present. In recent years, we have been pursuing a reverse evo-devo approach to understand the role of homeobox gene duplications, losses and sequence changes in evolution. Here we discuss the fates of duplicated homeobox genes, focusing on an underappreciated and important mode of evolution: asymmetric divergence after gene duplication.

2. *Cis*-regulatory evolution: not the only game in town

An over-simplification has crept into the field in evolutionary developmental biology. Consider the following three findings that have been widely commented on. First, the discovery, made over several years, that many genes used in development are highly conserved in sequence between disparate taxa (Hox genes, Pax genes, *hh* genes and many others). This led to the idea of a conserved 'genetic toolkit' for development, differing little between animal phyla [1–3]. Second, there have been several attention-grabbing demonstrations that genes from one species can partially mimic the phenotypic effects of those from another species in transgenic experiments (such as the classic experiment of ectopic mouse *Pax6* driving formation of eyes in *Drosophila* [4]). These experiments reveal trans-phyletic conservation of biochemical or cellular function, but they have also been used to give further weight to the idea of a universal toolkit, and hint that important evolutionary changes may not lie within the coding sequences of genes. Third, genetic association methods have been used to trace the molecular basis of small phenotypic differences between or within species, such as fin spine prominence in sticklebacks [5] or trichome density on *Drosophila* legs [6], and in several cases these have been traced to *cis*-regulatory changes. The modularity of *cis*-regulation, whereby one aspect of expression can be tweaked without affecting other aspects, is key. Together these findings have highlighted the importance of mutations affecting expression of genes, rather than the number of genes or their encoded amino acid sequences. A further issue that compounded this view is that before the advent of high-throughput transcriptomics and genomics, the dominant techniques for finding genes of interest were biased: methods such as PCR and low stringency library screening inevitably led to a focus on genes that are conserved between species.

We do not dispute the importance of these findings, and indeed we consider them among the most significant discoveries in the history of biology. The issue centres on the extent to which other sorts of mutation also play a role. In 2000, Carroll [7, p. 578] cited data in support of the claim that 'regulatory DNA is the predominant source of the genetic diversity that underlies morphological variation and evolution', and similarly in 2008 argued that 'form evolves largely by altering the expression of functionally conserved proteins, and . . . such changes largely occur through mutations in the *cis*-regulatory sequences of pleiotropic developmental regulatory loci and of the target genes within the vast networks they control' [1, p. 25]. We suggest that the words 'predominant' and 'largely' cannot yet be justified, as we do not have a quantitative assessment of the relative roles played by different sorts of mutation across animal evolution. Others have made the same point, and indeed Hoekstra & Coyne [8, p. 995] stated 'Although this claim may be true, it is at best premature'. The problem is that the claim of *cis*-regulatory primacy can be misinterpreted to suggest that *cis*-regulatory change is all that needs to be considered for the evolution of form. This focus does a disservice to the field of evolutionary developmental biology as many other forms of mutation have occurred in evolution and we need to understand their significance.

There are several ways in which protein-coding sequences can (and do) evolve, and which are relevant to the evolution of developmental mechanisms. Examples are known of deletion of genes causing phenotypic change (e.g. in the artificially selected twin-tailed goldfish [9]) and many cases of small mutational changes in coding sequences causing subtle changes to the DNA-binding specificity of the encoded transcription factors, protein stability or cofactor interactions (reviewed by [10]). These sorts of mutations can occur in the absence of gene duplication or in one paralogous gene after duplication. Here we focus on such changes after gene duplication, and especially cases where one or more paralogues accumulate radical change while a sister gene remains essentially unchanged.

3. Gene numbers: up and down in evolution

The number of genes present in the genome varies by several thousand between animal species. Deducing the precise number of protein-coding genes in a genome is extremely hard, even with a 'complete' genome sequence, because of difficulties in recognizing short protein-coding genes, distinguishing functional genes from non-functional pseudogenes, and assembling chromosomal regions containing repeats and duplications. Nonetheless, even within a group of relatively closely related species, such as placental mammals, the numbers of predicted genes varies by several thousand. For example, the US National Center for Biotechnology Information (NCBI) currently lists total protein-coding gene numbers to be: human 20 254; mouse 22 504; dog 19 871; cow 21 498 (www.ncbi.nlm.nih.gov/gene/; accessed 10 May 2016) (figure 1). Moving to animals outside the mammals reveals even more differences between species; for example, *Ciona intestinalis* 13 648; *Drosophila melanogaster* 13 919; *Caenorhabditis elegans* 20 269.

What is the basis for these numerical differences between taxa? The differences reflect additions of genes and losses of genes. Change in total gene number represents a net balance between gain and loss, and hence the true rate of gene gain in evolution must be greater than suggested by the raw numbers alone. Indeed, most published genome projects typically find hundreds or thousands of genes with no clear orthologues in other species, indicative of evolutionarily recent lineage-specific change. There are several routes to gaining genes, including whole-genome duplication (WGD), tandem gene duplication (TGD), segmental duplication (essentially giant multi-gene tandem duplication), retroposition and complex combinations of exon copying, de novo incorporation of non-coding DNA and fusion of mobile genetic elements. Recent estimates suggest that less than 1% of human genes arose by retrotransposition: approximately 160 retrogenes have parental copies still existing plus for approximately 25 'orphan' retrogenes the parental gene has been lost [11,12].

There are several well-documented cases of WGD in animal evolution, and although these had an impact in some evolutionary lineages, they are not the main reason

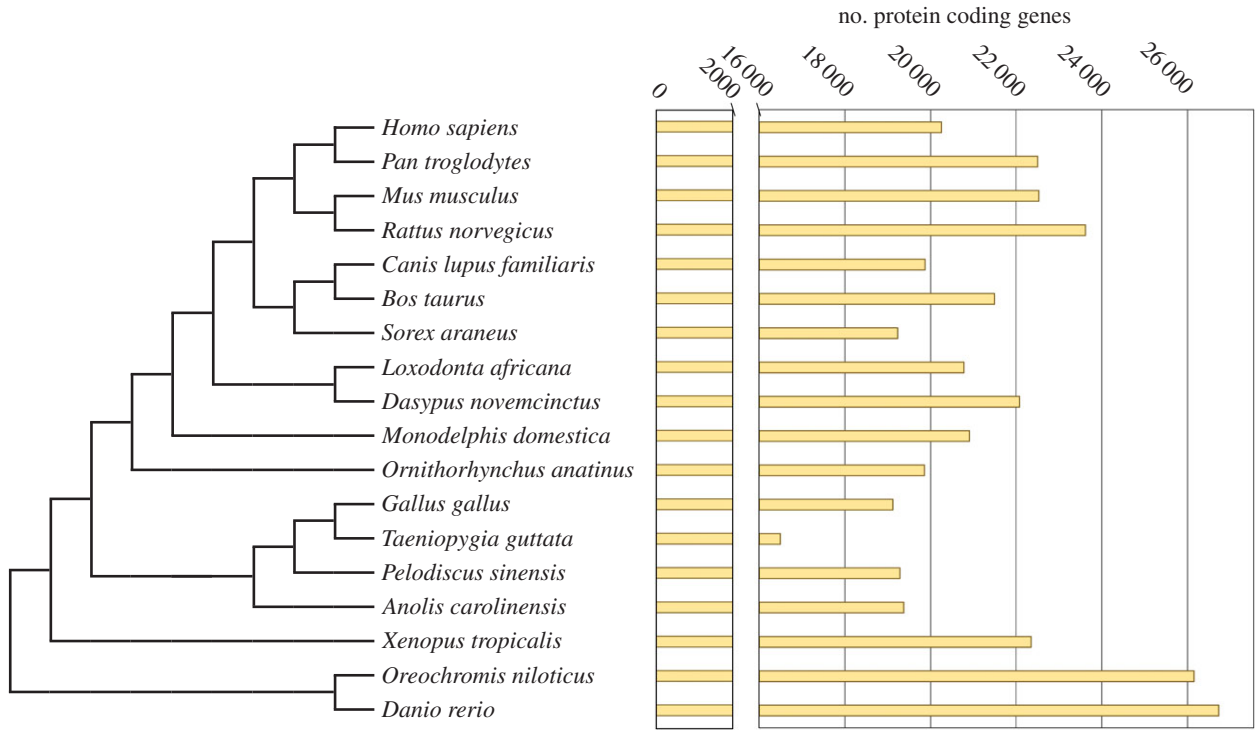

Figure 1. Histogram showing numbers of protein-coding genes predicted in the genomes of a range of vertebrates, compared with phylogenetic relationships. Dynamic gene gain and loss along each lineage causes variation in the range of hundreds to thousands of genes between related species. Protein-coding gene numbers from NCBI (www.ncbi.nlm.nih.gov/genes; accessed 10 May 2016) using taxon identification number (Txid) and specifying Reference Sequence (RefSeq) genes only. (Online version in colour.)

for the gene number differences. WGD occurred twice at the base of vertebrates [13,14], once in the stem lineage of teleost fish [15–17], one or more times in cyprinid fish [18], once in salmonid fish [19], two or more times in chelicerates [20,21] and once in rotifers [22]; undoubtedly more cases will be discovered. However, a clear finding is that gene loss is extensive after WGD, possibly reaching as high as 85% of duplicates in some cases [23]. This is not to say that WGD is unimportant in animal evolution, far from it. There is a general rule of vertebrates having more genes than invertebrates (though there are exceptions), which is likely to be traceable to WGD, and from a functional perspective it is relevant that a comparison of amphioxus and vertebrate genomes revealed that genes encoding transcription factors or deployed in development and neural function are among those retained preferentially after WGD [24]. The implication is that while WGD may not have caused massive changes in total gene number, it expanded small subsets of developmentally important genes that could be recruited to new roles [23].

To return to the question of total gene number differences, we consider a prevalent source of gene number differences to be TGD. There are many well-characterized cases of expansion of particular sets of genes in certain lineages, associated with changes in physiological, structural or behavioural traits. For example, honeybees have unusually large numbers of odorant receptor genes [25], dragonflies have a large expansion of opsin genes associated with high-acuity vision [26] and oysters have large numbers of genes for heat shock proteins expressed during stress at low tide [27]. Among cephalopods, *Octopus* has an expansion of genes coding for protocadherins and also genes encoding C2H2 zinc finger transcription factors, which are both involved in neural development and patterning [28]. In each of these cases, the duplicated genes have undergone rather subtle evolutionary divergence from each other. There are also many cases of gene family expansion for which the underlying adaptive reasons are unclear. For example, there was expansion of KRAB-box zinc finger genes in mammalian evolution [29] and the *Caenorhabditis* genus of nematodes has large numbers of nuclear hormone genes [30]. In this paper, we cannot review the many cases of gene duplication across the animal kingdom, but will focus on the homeobox gene superclass.

4. Asymmetric evolution: an underappreciated route to novelty

Before discussing examples, it is necessary to understand why identifying duplicated genes can sometimes be difficult. Unless we can identify which genes have duplicated, we cannot make statements about their contribution to developmental evolution. When a gene duplicates, two loci are generated (more than two could be generated, but simpler to consider a mutation that generates two loci from one). It is formally incorrect to denote one locus as the 'parent' or 'original' gene, and the other as the 'duplicate' or 'daughter' gene. One is not 'old' and one is not 'new'; both loci are the same age. The two loci are equally orthologous to the single gene retained in a sister taxon without duplication, and indeed the terms semi-orthologue and pro-orthologue were coined to describe this many-to-one homology relationship [31,32]. The situation is not so simple if, for example, the mutation copies only part of a locus. This could well be the situation for many cases of TGD, where only some of

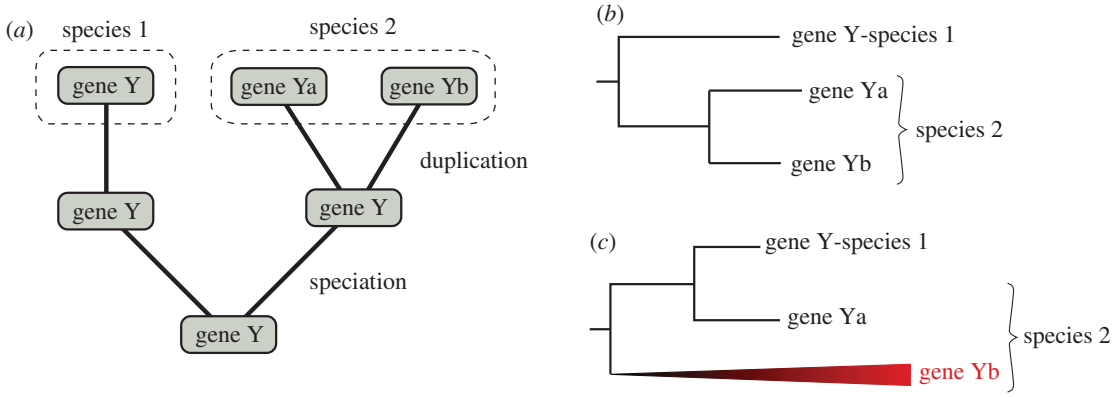

Figure 2. Asymmetric evolution of gene duplicates. (*a*) Gene Y duplicates in species 2 but not in species 1. (*b*) If the duplicate genes (Ya and Yb) accumulate sequence change at approximately equal rates, phylogenetic reconstruction methodologies will readily recover the correct evolutionary history. (*c*) If gene Yb accumulates sequence changes at a far greater rate than gene Ya or the pro-orthologue in species 1, this constitutes asymmetric evolution. Phylogenetic reconstruction may fail to recover the true evolutionary history, and can mistakenly place gene Yb as an outgroup of Ya and Y. (Online version in colour.)

the regulatory landscape is duplicated, but the simple situation will always be the case for WGD events when the entire locus is precisely duplicated. Leaving aside this complication for the moment, the normal situation would be that duplication can generate two loci of identical age, sequence, expression and function.

The duplicate loci will each accumulate mutations over time, in coding sequences and regulatory elements, of a neutral, deleterious or adaptive nature. They are unlikely to accumulate the same mutations. In many cases, we might expect both duplicates to diverge from the ancestral state at a similar rate. We denote this mode of evolution 'symmetrical' divergence (figure 2*a*,*b*). The term symmetrical is not used in a strict sense of absolute equity, but only to denote approximately equivalent amount of change. For example, if the four Hox gene clusters of mammals are compared with the single Hox gene cluster of amphioxus, all are approximately equidistant in encoded amino acid sequence. In addition, all retain the inferred ancestral expression in central nervous system, complemented by subtle differences in deployment to other tissues such as somatic and visceral mesoderm. Similarly, the three *Cdx* genes of *Xenopus tropicalis* are roughly equidistant from the single *Cdx* gene of amphioxus, and functional interference indicates subtle differences in developmental roles [33,34]. More cases could be given including the two vertebrate *En* genes, the two or three *Emx* genes, the two *Gsx* genes, the three *Tlx* genes, etc. In all these cases, the duplicates derive from a WGD, and the same symmetrical pattern is seen for almost all vertebrate WGD-derived homeobox gene duplicates. Symmetrical divergence is not confined to WGD, but can also be seen for many tandemly duplicated loci, such as *AmphiEmxa* and *AmphiEmxb* in amphioxus [35], although in some cases gene conversion between tandem duplicates can reinforce the similarity after divergence (as in many cases in insects, such as *engrailed* and *invected* [36]).

In contrast with symmetrical patterns of divergence, several cases of strikingly unequal divergence of duplicates have been found where one locus changes radically from the ancestral sequence and the other locus changes relatively little. For example, Steinke and colleagues compared genomes between three teleost fish, plus human as an outgroup and identified many cases of duplicate gene pairs in which one paralogue had experienced a lineage-specific elevated rate of molecular evolution [37]. This is denoted asymmetric divergence. There are three important consequences of asymmetric divergence. First, there is a nomenclature problem in that there is a temptation to consider the radically changed locus as a 'daughter' gene and the locus that changed little as a 'parent' or 'original' genes, even though technically the two loci are the same age. We will use the parent/daughter terminology as it is often pragmatic to do so, while being mindful of its limitations. Second, rapid divergence of one daughter gene gives great potential for recruitment to new biological roles, and this is likely to have profound consequences for developmental evolution. Instead of subfunctionalization of ancestral roles or addition of extra roles, completely new roles could evolve. Third, there is a practical problem that if one locus diverges greatly, it can sometimes be very difficult to deduce how it has arisen, because an accelerated evolutionary rate can induce phylogenetic reconstruction artefacts such as 'long branch attraction' [38,39]. Indeed, there are cases where the divergence is so great that in phylogenetic trees the duplicate is misplaced and erroneously appears as an outgroup to the unduplicated gene and its sister gene (figure 2*c*). We consider this to be a serious problem and one that has contributed to an underappreciation of the importance of asymmetric evolution. Seemingly 'new' genes are found in every genome analysed, including 'lineage-specific' homeobox genes; most of these will have arisen by asymmetric divergence but the precise pathway of evolution is often unknown.

Asymmetric divergence has been found in homeobox gene families generated by the vertebrate WGD events, but it does not seem to be common after WGD. For example, when homologues of the *Drosophila* gene *orthodenticle* (*otd*) were first described in mammals, only two were found—denoted *Otx1* and *Otx2* [40]. This was in the earlier days of homeobox gene comparisons between species, and there was much excitement in the finding that the two mammalian genes are expressed in the developing head and brain, in a comparable way to *Drosophila otd* [40,41]. Several years later it was recognized that there is a third member of the *Otx* gene family in vertebrates, a divergent gene with quite different expression to *Otx1*, *Otx2* or *otd*, or indeed to amphioxus *Otx*: a gene that is retinal-expressed in mammals and denoted *Crx* (*Cone-rod homeobox*). Consideration of chromosomal position reveals that *Otx1*, *Otx2* and *Crx* were generated by the WGD events, but *Crx* has diverged most from the ancestral sequence and expression pattern [42]. A second example concerns the much-studied *Pax6* gene, known to be

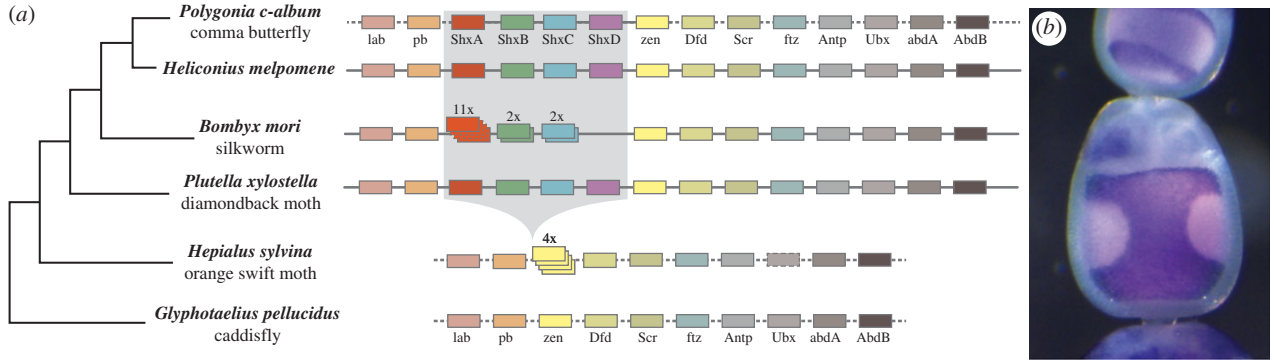

Figure 3. Asymmetric evolution of *zen* gene duplicates in Lepidoptera. (*a*) The *zen* gene (Hox paralogy group 3 gene) of insects duplicated to give four additional Hox cluster genes, *ShxA* to *ShxD*, within the Lepidoptera; these underwent extensive sequence divergence. The basal Orange swift moth also has *zen* duplicates, but without extensive divergence. (*b*) Localized *ShxC* RNA marks the presumptive serosa in developing oocytes of the speckled wood butterfly *Pararge aegeria* [52]. All four *Shx* genes are expressed in serosa as it develops.

orthologous to *Drosophila eyeless*, with both genes playing important roles in eye development [43]. Despite the massive amount of work on *Pax6*, it was only relatively recently that it was confirmed that vertebrates have a WGD-derived paralogue of this gene: a previously known gene called *Pax4* [44]. Again, chromosomal position confirms that *Pax6* and *Pax4* are of equal age, both generated by the vertebrate WGD events. Thus while *Pax6* has diverged relatively little, *Pax4* has changed radically in sequence and regulation, and now plays a role in vertebrate pancreas development [45]. In this case, the sequence divergence is so large that phylogenetic analysis places *Pax4* erroneously: it looks like a 'novel' gene without a close invertebrate homologue, whereas in reality it is simply the paralogue of *Pax6* but the extent of sequence divergence has caused violation of the assumptions of phylogenetic inference programmes.

Asymmetric divergence may be relatively rare in homeobox genes after WGD (with *Oxt1*/*Otx2*/*Crx* and *Pax4*/*Pax6* being important exceptions), but we find it is common after TGD. Below we describe several cases where asymmetric divergence has generated 'novel' homeobox genes that have been recruited for new developmental roles.

5. Extra Hox genes and the evolutionary success of Lepidoptera

Most of the described species of animals are insects; indeed, it has been noted that 'to a good approximation, all species are insects' [46, p. 514]. Within the insects, the 'big four' orders are Coleoptera (beetles), Diptera (flies), Hymenoptera (bees, wasps, ants, etc.) and Lepidoptera (butterflies and moths). There are around 150 000 described species of Lepidoptera; three times more than all vertebrates, for example. It may be futile to search for simple explanations for why there are so many butterflies and moth species, but some contributing factors can be postulated. In particular, butterflies and moths most likely radiated in concert with the diversification of flowering plants [47], with the larvae of each species evolving adaptations to allow phytophagy on (or in) leaves, usually in the face of intense chemical defence from the plants [48]. Lepidoptera have evolved sophisticated and adaptable detoxification systems to overcome such defences [49,50]. But chemical attack is not the only barrier to leaf-feeding. Many Lepidoptera lay their eggs on the surface of leaves leaving them exposed to the dangers of desiccation and attack by fungi and bacteria; those species that lay eggs inside leaves will face similar threats. Dropping eggs onto damp soil leading to subterranean root-feeding larvae, a strategy used by the basal Hepialidae, at least overcomes the desiccation threat. Perhaps surprisingly, tandem duplication and asymmetric divergence of Hox genes may have partly contributed to desiccation protection and immune defence in the eggs of Lepidoptera.

The Hox gene cluster is highly conserved across insects. It was, therefore, exciting when Chai *et al.* [51] reported that the Hox gene cluster of the domesticated silkmoth *Bombyx mori* contains at least 11 highly divergent homeobox loci, additional to the expected *Hox* genes, located between *pb* and *zen*. Subsequent analysis revealed the number is most probably 15 [52]. The *Bombyx Hox* cluster is still the largest known in any animal, in terms of gene number, making the discovery by Chai *et al.* [51] a very significant one in the field of evolutionary genomics. It was not known at the time how many of these genes, termed *Shx* (Special Homeobox) genes, are functional and certainly some Shx loci have mutations in the coding sequence. Evidence that these genes are not unique to *Bombyx* came from analysis of the genome of a butterfly *Heliconius melpomone*, found to have four *Shx* genes located between *pb* and *zen* [53].

To investigate the origin of Shx genes, we determined low-coverage genome sequences for five additional lepidopteran species chosen for phylogenetic position, plus a caddisfly outgroup [52,54]. Assembly and analysis revealed that the *Bombyx* situation is unusual, whereas possession of four distinct *Shx* genes (*ShxA*, *ShxB*, *ShxC*, *ShxD*) is typical for a large clade within Lepidoptera, encompassing butterflies (*Heliconius*, *Polygonium*, *Pararge*), tiger moth (*Callimorpha*) and Gracillariidae (*Cameraria*). Together these lineages fall within the Ditrysia, sometimes referred to as 'higher Lepidoptera'. By contrast, we found that the caddisfly outgroup and a representative of the basal family Hepialidae (*Hepialus*) lack Shx genes (figure 3*a*).

The homeodomain sequences of *Shx* genes are very different from the canonical Hox genes, with long branch lengths on phylogenetic trees implying rapid sequence change in evolution. If the *Shx* genes had been found located in a different genomic region, it may have been difficult to ascertain their origin with certainty. However, their location precisely between *pb* and *zen* strongly indicates that these genes

arose by tandem duplication from a *Hox* gene, followed by extensive sequence divergence. Indeed, phylogenetic analysis places *Shx* genes as a sister group to *zen*, albeit with long branches, revealing that *zen* is almost certainty the 'parental' gene from which *Shx* genes arose. It is important to stress that the *zen* gene is still present in all the Lepidoptera species possessing *Shx* genes. The implication is that the *zen* gene duplicated to give two (or more) identical copies, followed by extensive sequence divergence in the duplicates lying closest to *pb*. By contrast, the *zen* copy closest to *Dfd* diverged far less and has remained a bona fide *zen* gene. The *zen*/*Shx* genes, therefore, constitute a very clear case of asymmetric sequence divergence. It is possible that the Hepialidae share the same duplication but not the same pattern of sequence divergence, because although *Hepialus* does not possess *Shx* genes it has multiple *zen* loci.

The sequence divergence of *Shx* genes, plus the maintenance of number between several species, suggests they have been recruited to new and conserved roles in Ditrysia. To investigate probable roles, our colleagues Casper Breuker, Jean-Michel Carter and Melanie Gibbs analysed expression patterns in the speckled wood butterfly, *Pararge aegeria* [52]. This revealed clearly localized expression of all four *Shx* genes in cells fated to become serosa, an important extra-embryonic membrane wrapping around the developing embryo. For *ShxA* and *ShxC*, there is also maternal RNA in the egg, which for *ShxC* is strikingly localized in a complex horseshoe shape within the unfertilized oocyte, marking the territory fated to become serosa (figure 3*b*). The serosa plays a critical role in defence of the egg against desiccation and pathogens, and is important to survival of eggs laid on exposed surfaces of vegetation as is common for most Ditrysia. We suggest, therefore, that Hox TGD and asymmetric divergence generated a novel set of homeodomain transcription factors that were recruited for specifying and patterning the serosa; this was one of the myriad of adaptations permitting success of the Lepidoptera.

6. Extra TALE-class genes and the development of molluscs

In the case of Shx genes, their location within the Hox cluster provided an important clue to the origin of the genes. Often the situation is not so clear because inversions and translocations can separate tandemly duplicated genes and leave them dispersed around the genome. In analysing the genome of the Pacific oyster *Crassostrea gigas* [27], we identified an expansion in the number of homeobox genes compared with the inferred ancestral number for bilaterian animals [55]. Fourteen of the novel genes have a characteristic TALE homeobox (encoding a homeodomain with a three amino acid insertion) and hence must have arisen by duplication and divergence from other TALE-class genes, although we have not identified the precise parental genes and define them as cryptic paralogues. We also identified nine PRD-class genes of uncertain origin [55]. In addition, the well-known gene *engrailed* also has two copies in lophotrochozoan lineages that show the signature of asymmetrical evolution [55]. During development of the oyster, *en2* displays a peak of expression in the gastrula stage and is expressed in the mantle of adults where it has been implicated in formation of the shell characteristic of molluscs [27]; by contrast, *en1* shows more homogeneous expression levels across tissues and stages. This illustrates the acquisition of divergent functions during asymmetric evolution.

Comparison with genome data from seven other lophotrochozoan genomes revealed that the additional TALE and PRD homeobox genes arose at various times in the ancestry of oysters, with some shared across lophotrochozoans, some shared between annelids and molluscs, several found only in molluscs and a few restricted to bivalves. Analysis of extensive transcriptome data revealed that most of the new genes (especially TALE-class genes) have peak expression at very early developmental stages, during cleavage and blastula formation, while several others (mostly PRD genes) are most highly expressed much later after the trochophore stage (figure 4). The recruitment of new homeobox genes, or rather highly diverged duplicated homeobox genes, to early and late development is intriguing. First, there is an interesting parallel to the Shx example above, as in both cases, the new homeobox genes have been recruited to very early embryonic stages when initial cell fate decisions are made. Second, the patterns are consistent with the much discussed hourglass or egg-timer model for the evolution of development, which postulates that early and late stages of embryonic development are less constrained and more able to tolerate modification in evolution [56]. It seems that this modification can involve incorporation of new divergent transcription factors into the regulatory landscape, highlighting the importance of asymmetric evolution in remodelling gene regulatory networks.

7. Extra PRD-class genes and the development of placental mammals

An example of asymmetric evolution that is gaining considerable attention concerns a set of homeobox genes in the genome of humans and other eutherian (placental) mammals. After the initial drafts of the human genome were released in 2001, we set out to identify, annotate and classify all human homeobox genes. This survey has been progressively revised [57–60]. One of the most interesting findings was our discovery of several novel PRD-class homeobox genes, including five which we named *ARGFX*, *TPRX1*, *TPRX2* (initially called *TPRX2P*), *DPRX* and *LEUTX* [57,59]. These genes were previously undescribed, unnamed and without clear orthologues in mouse or other animal genomes characterized at the time. Other human PRD-class genes with similarly restricted distributions include *CPHX1*, *CPHX2* and the double homeobox genes *DUXA* and *DUXB* [57,60–62] (table 1). The *TPRX1* and *TPRX2* genes flank the Otx family gene *CRX* at 19q13, suggestive of origin by tandem duplication and divergence, and several of the other genes including *LEUTX* and *DPRX* are more distant on the long arm of chromosome 19. An origin from *CRX* for most of these genes has been proposed, and recently we have confirmed this for *ARGFX*, *DPRX*, *TPRX1*, *TPRX2* and *LEUTX* using a combination of molecular phylogenetic analysis and examination of conserved non-coding elements [61,63] (figure 5*a*).

The phylogenetic distribution of the genes and their newly acquired developmental roles are of particular interest. While the DUX genes have a probable orthologue outside eutherian mammals (with one rather than two homeoboxes; [62]), the other genes are restricted to eutherians. Not every

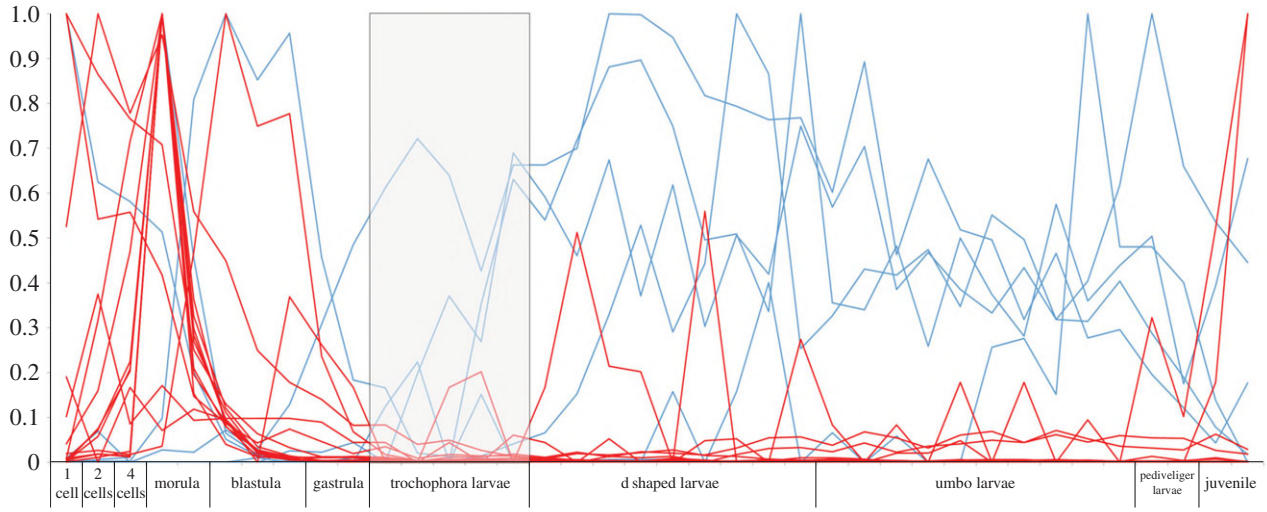

Figure 4. Temporal expression patterns of novel PRD (blue) and TALE (red) genes in the Pacific oyster across 38 developmental stages (data from [55]). Expression levels were normalized for each gene, with the value 1 being the maximum expression level of a gene across the temporal series. All the genes peak in developmental stages before or after the trochophore larva phase (indicated by the grey box).

Table 1. Diversity, location and origin of new mammalian PRD-class homeobox genes.

new PRD gene family	functional genes in human	human chromosome (s)	no. pseudogenes	parental gene
Argfx	*ARGFX*	3	2	Crx
Dprx	*DPRX*	19	7	Crx
Leutx	*LEUTX*	19	0	Crx
Tprx	*TPRX1*, *TPRX2*	19	3	Crx
Pargfx	—	—	1	Crx
Dux	*DUXA*, *DUXB*	19,16	approximately 34 (some may be functional)	sDux
Cphx	*CPHX1*, *CHPX2*	16	2	not known

gene is found in every eutherian species because of gene losses; for example, one gene (*Pargfx*) found in horse and dog has been lost in humans, and many of the genes are lost in mice and rats [61,63]. In summary, it can be deduced that the *Argfx*, *Dprx*, *Tprx* and *Leutx* gene families arose by tandem duplication from the *Crx* gene in the stem lineage of the placental mammals, after their split from marsupials (figure 5*a*). These 'new' PRD-class then underwent radical divergence, while *Crx* evolved slowly, in a very clear case of asymmetric evolution. The extent of sequence divergence from the 'parental' gene, from each other and from orthologues from different mammalian species is extremely high; for example, human *LEUTX* shares only 29% identity with *CRX* over the alignable region, and only 42% in the homeodomain. As a comparison, the homeodomains of the three human Otx genes (*OTX1*, *OTX2*, *CRX*), which diverged more than 450 Ma after the two rounds of WGDs, are 87% identical (52 in 60). Further, the Crx homeodomain shares approximately 60% identity with that of a non-Otx PRD-class gene homeobox Gsc, a gene that diverged from Otx genes before the origin of bilateral animals. Thus, during approximately 100 Ma since the origin of eutherians, *LEUTX* has diverged more from its parental gene *CRX*, than *CRX* has diverged from other PRD homeoboxes since the Cambrian explosion (figure 5*c*).

Our first clue to possible functions of *ARGFX*, *TPRX* and *DPRX* genes came from the observation that these genes, plus *DUXA*, had generated processed pseudogenes in human evolutionary history. As this type of pseudogene derives from retrotransposition of mRNA, their presence in the inherited genome is a clear indication that the gene must be expressed in the germ line [57,58]. For example, several important genes expressed in the pluripotent cells of the blastocyst and embryonic stem cells, including *NANOG* and *POU5F1*, also have multiple processed pseudogenes [58,65,66]. In contrast with *NANOG* and *POU5F1*, we could not find expression of *ARGFX*, *DPRX*, *TRPX1*, *TPRX2* or *LEUTX* in human embryonic stem cells (hESC), nor could we find strong expression in another germ line tissue, the testis. These negative results were confusing, as they seemed at odds with the presence of processed pseudogenes. The conundrum was resolved when transcriptome data were published for the earliest stages of human development, before blastocyst formation [64,67], and therefore, earlier than the stages mimicked by hESC. These data revealed that human *ARGFX*, *DPRX*, *LEUTX*, *TPRX1* and *TPRX2* are specifically expressed from 8-cell to morula in a striking and specific pulse of expression just before cell fates are established [61,63] (figure 5*b*). These pre-blastocyst embryonic stages are totipotent and will form every tissue of the

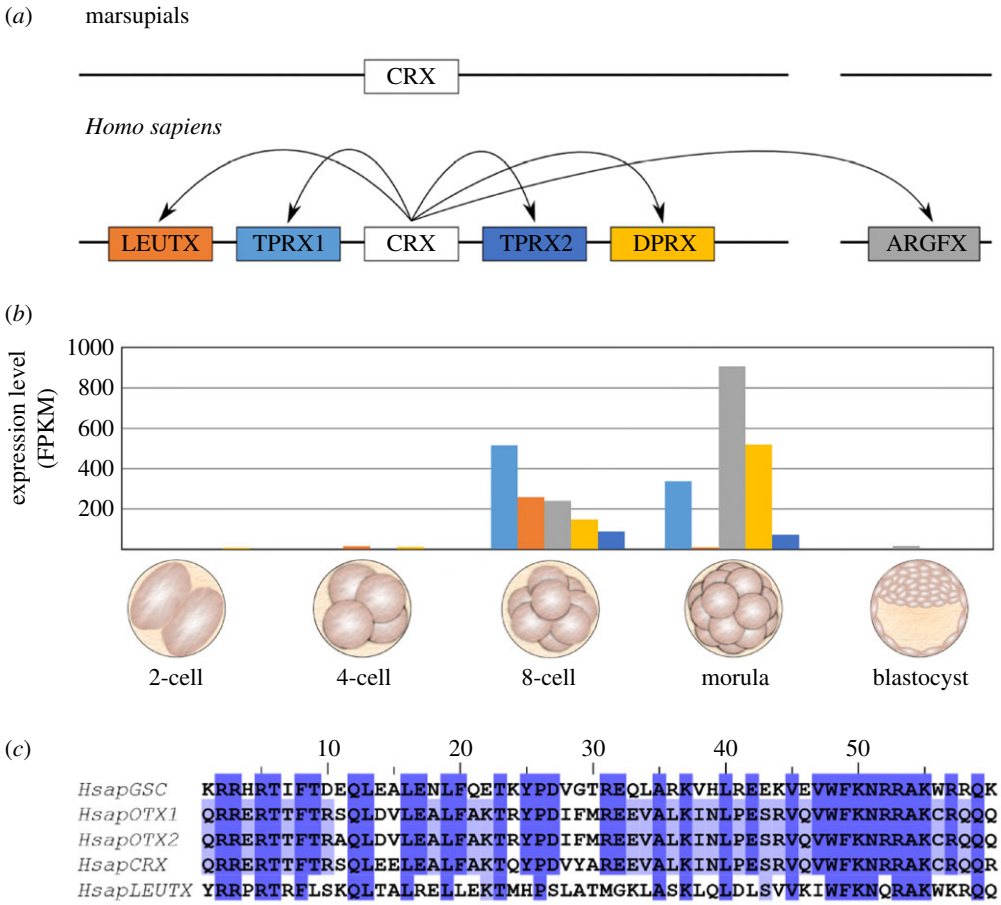

Figure 5. Asymmetric evolution of *Crx* gene duplicates. (*a*) The divergent PRD-class homeobox genes *Leutx*, *Tprx1*, *Tprx2*, *Dprx* and *Argfx* arose in placental mammals by tandem duplication from the *Crx* gene, a member of the Otx gene family. (*b*) The novel genes are expressed specifically from the 8-cell to morula stages of pre-implantation human development. FPKM (fragments per kilobase per million sequence reads) values obtained by RNAseq mapping reads [64] to human genome hg38 assembly [61,63]. Histogram bars, from left to right: *TPRX1*, *LEUTX*, *ARGFX*, *DPRX* and *TPRX2*. (*c*) Alignment of the homeodomains of human *LEUTX*, *CRX*, *OTX1*, *OTX2* and *GSC*. Identical residues shared by all genes shown are highlighted in dark blue and those shared by the three *OTX* proteins (*OTX1*, *OTX2*, *CRX*) in light blue. Note that just a small subset of the conserved Otx family residues are also conserved in *LEUTX*.

embryo and extra-embryonic membranes. Deciphering the roles of the genes in human pre-blastocyst stages is fraught with ethical and practical difficulties, and no cell line proxies exist. Two indirect approaches have been used. First, Töhönen *et al.* [61] used a bioinformatic approach to identify putative promoter motifs enriched at this stage of development, and showed enrichment in putative binding sites for PRD-class proteins. This suggested transcription factor roles for the divergent PRD-class proteins, and allowed some possible targets to be postulated. Second, both our laboratory and Madissoon *et al.* used a transfection approach to ectopically express several of the genes in human cells (fibroblasts or embryonic stem cells), followed by RNAseq to identify transcriptional changes [63,68]. In our analyses, we uncovered many downstream effects including activation and repression of a set of genes that have a similar 'pulse' of expression in the human morula [63]. The conclusion is that these newly arisen genes have been recruited for very specific roles in the earliest developmental stages of placental mammals, including the human embryo.

In summary, the PRD-class genes provide another clear case of origin by tandem duplication and asymmetric sequence evolution leading to the evolution of 'new' homeobox genes. In this case, the genes arose in the stem lineage of the eutherian mammals and were recruited for novel developmental functions at precisely the time when the distinction between embryonic and extra-embryonic tissues is being established. Such tissues are, of course, vital for placental development.

8. Future perspectives

In this article, we have explained the nature of asymmetric evolution and provided examples showing how this process has resulted in 'new' genes that were recruited for new roles in development. To evaluate the broader significance of this process for the evolution of diversity, and morphological evolution in general, it will be necessary to deduce how widespread this mode of molecular evolution has been. As noted above, asymmetric evolution is best characterized in connection with TGD, and fewer examples are known following WGD in vertebrates. One future task will be to rigorously examine whether this difference is a general rule, albeit one with exceptions. There may be a good reason why asymmetric evolution is commoner after tandem duplication, which is that tandem duplication will always disrupt the genomic environment of genes, for example, by copying only part of the regulatory landscape. This, in turn, could predispose one duplicate copy to diverge functionally. By contrast, WGD genuinely results in identical gene copies.

If we accept that TGD is more likely to lead to asymmetric gene evolution, then a second major task will be to deduce the prevalence of tandem duplication. Has there been a slow and steady 'drip feed' of TGD in animal evolution or more of a flood? There are two reasons why simply counting the number of duplicated genes underestimates the true rate of tandem duplication.

The first reason is gene loss. This can be seen even among highly conserved and ancient homeobox genes, where almost all evolutionary lineages have experienced some loss. Examples of genes that were present in the genome of the first bilaterian but were later lost in different lineages include *Barx* (lost in Ecdysozoa), *Hopx* (so far only found in chordates and molluscs) and *Pou1* (lost in the ancestor of ecdysozoans, and in multiple lophotrochozoans) [55]. Tapeworms are an example where whole-genome sequencing revealed extensive loss of homeobox genes [69]. As parasites, tapeworms have many specializations such as a complex tegument used for absorption of nutrients from the host. But they have also lost many features, including the gut, a complex brain, eyes and muscles used for motility. This is in addition to other simplifications seen in all Platyhelminthes. When the genomes of four tapeworm species were sequenced, it was possible to ask whether any developmentally important genes, such as those encoding transcription factors or signalling molecules, had been lost in evolution. We found that tapeworms have lost around a third of all homeobox genes generally present in bilaterian animals; instead of 96 homeobox genes that date to the base of Bilateria, tapeworms have only approximately 62 (with losses occurring at different times in evolution; [69]). It cannot be concluded that the disabling mutations responsible for gene loss actually caused developmental changes and loss of structures; however, the loss of over 30 (otherwise conserved and essential) homeobox genes is one of the most striking examples of co-evolution of genomes and morphology. The above example concerns homeobox genes that have been highly conserved across other animals. Gene loss is likely to have a higher probability for genes that are less conserved, and thus in general counting duplicated genes is likely to always underestimate gene duplication rate, because of occasional gene loss.

The second reason that gene duplication is underestimated is less obvious, and relates to the dynamic nature of the genome. When examining the genomic regions around the *Crx*-derived mammalian PRD-class genes discussed above, we found much evidence for dynamic gain and loss of genes. For example, a conserved non-coding element associated with the parental and daughter genes is also found in additional copies, with no neighbouring gene, implying that additional *Crx*-derived genes had been generated, but then lost from all extant lineages [61,63]. In addition, many cases of recent tandem duplication are observed that must have occurred subsequent to the origin of the genes, including additional *Tprx* loci in tenrec, bat, horse, mouse and rat, and additional *Leutx* genes in guinea pig and elephant [61,63]. Gene loss is also prevalent. The picture is one of a genomic region that is generating new loci and losing loci at a high rate. In any situation with rapid gain and loss of characters across multiple lineages, the principle of parsimony breaks down. This has been amply demonstrated for DNA and protein sequence change [39] and is equally relevant for changes to numbers of genes. The simple assumption that the gene numbers in each species were generated by the evolutionary pathway that involved the fewest gain and loss events is almost certainly wrong. It is much more likely that a model of continuous gain and loss, or 'gene turnover', gave rise to the observed pattern. The implication is that every species has lost genes that are no longer observed, and the rate of TGD is even higher than is suggested by counting gene numbers. We suggest this principle will not be unique to the *Crx* chromosomal region, but will extend to many cases of TGD. We suggest, therefore, that the gain of genes by TGD is more of a flood than a slow drip. Many of the genes generated are rapidly lost, leaving a fraction to be captured for novel roles by natural selection, often through asymmetric divergence.

Authors' contributions. All authors contributed to drafting and revising the article.

Competing interests. We have no competing interests.

Funding. This work was supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013 ERC grant 268513).

Acknowledgement. We thank Casper Breuker, Jean-Michel Carter, Nathalie Feiner, Laura Ferguson, Adam Hargreaves, Jerome Hui, Gil McVean, Shan Quah, Amy Royall, Sebastian Shimeld and Fei Xu for stimulating discussions.

References

1. Carroll SB. 2008 Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36. (doi:10.1016/j.cell.2008.06.030)
2. Rokas A. 2008 The origins of multicellularity and the early history of the genetic toolkit for animal development. *Annu. Rev. Genet.* **42**, 235–251. (doi:10.1146/annurev.genet.42.110807.091513)
3. Ingham PW, Nakano Y, Seger C. 2011 Mechanisms and functions of Hedgehog signalling across the metazoa. *Nat. Rev. Genet.* **12**, 393–406. (doi:10.1038/nrg2984)
4. Halder G, Callaerts P, Gehring WJ. 1995 Induction of ectopic eyes by targeted expression of the eyeless gene in Drosophila. *Science* **267**, 1788–1792. (doi:10.1126/science.7892602)
5. Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jónsson B, Schluter D, Kingsley DM. 2004 Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717–723. (doi:10.1038/nature02415)
6. Stern DL. 1998 The *Ubx* alleles from *D. melanogaster* and *D. simulans* produce different trichome patterns in identical hybrid backgrounds. *Nature* **396**, 463–466. (doi:10.1038/24863)
7. Carroll SB. 2000 Endless forms: the evolution of gene regulation and morphological diversity. *Cell* **101**, 577–580. (doi:10.1016/S0092-8674(00)80868-5)
8. Hoekstra HE, Coyne JA. 2007 The locus of evolution: evo devo and the genetics of adaptation. *Evolution* **61**, 995–1016. (doi:10.1111/j.1558-5646.2007.00105.x)
9. Abe G, Lee S-H, Chang M, Liu S-C, Tsai H-Y, Ota KG. 2014 The origin of the bifurcated axial skeletal system in the twin-tail goldfish. *Nat. Commun.* **5**, 1–7. (doi:10.1038/ncomms4360)
10. Jarvela AMC, Hinman VF. 2015 Evolution of transcription factor function as a mechanism for changing metazoan developmental gene regulatory networks. *EvoDevo* **6**, 3. (doi:10.1186/2041-9139-6-3)

11. Ciomborowska J, Rosikiewicz W, Szklarczyk D, Makałowski W, Makałowska I. 2013 'Orphan' retrogenes in the human genome. *Mol. Biol. Evol.* **30**, 384–396. (doi:10.1093/molbev/mss235)
12. Pan D, Zhang L. 2009 Burst of young retrogenes and independent retrogene formation in mammals. *PLoS ONE* **4**, e5040. (doi:10.1371/journal.pone.0005040)
13. Holland PWH, Garcia Fernàndez J, Williams NA, Sidow A. 1994 Gene duplications and the origins of vertebrate development. *Development* **1994**, 125–133.
14. Dehal P, Boore JL. 2005 Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*. **3**, e314. (doi:10.1371/journal.pbio.0030314)
15. Amores A *et al.* 1998 Zebrafish *hox* clusters and vertebrate genome evolution. *Science* **282**, 1711–1714. (doi:10.1126/science.282.5394.1711)
16. Jaillon O *et al.* 2004 Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**, 946–957. (doi:10.1038/nature03025)
17. Meyer A, van de Peer Y. 2005 From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *BioEssays* **27**, 937–945. (doi:10.1002/bies.20293)
18. Xu P *et al.* 2014 Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. *Nat. Genet.* **46**, 1212–1219. (doi:10.1038/ng.3098)
19. Macqueen DJ, Johnston IA. 2014 A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc. R. Soc. B* **281**, 20132881. (doi:10.1098/rspb.2013.2881)
20. Kenny NJ *et al.* 2015 Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. *Heredity* **116**, 190–199. (doi:10.1038/hdy.2015.89)
21. Nossa CW, Havlak P, Yue J-X, Lv J, Vincent KY, Brockmann HJ, Putnam NH. 2014 Joint assembly and genetic mapping of the Atlantic horseshoe crab genome reveals ancient whole genome duplication. *Gigascience* **3**, 708–721. (doi:10.1186/2047-217X-3-9)
22. Flot J-F *et al.* 2013 Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature* **500**, 453–457. (doi:10.1038/nature12326)
23. Brunet FG, Roest Crollius H, Paris M, Aury J-M, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M. 2006 Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol*. **23**, 1808–1816. (doi:10.1093/molbev/msl049)
24. Putnam NH *et al.* 2008 The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071. (doi:10.1038/nature06967)
25. Robertson HM, Wanner KW. 2006 The chemoreceptor superfamily in the honey bee, *Apis mellifera*: expansion of the odorant, but not gustatory, receptor family. *Genome Res.* **16**, 1395–1403. (doi:10.1101/gr.5057506)
26. Futahashi R, Kawahara-Miki R, Kinoshita M, Yoshitake K, Yajima S, Arikawa K, Fukatsu T. 2015 Extraordinary diversity of visual opsin genes in dragonflies. *Proc. Natl Acad. Sci. USA* **112**, E1247–E1256. (doi:10.1073/pnas.1424670112)
27. Zhang G *et al.* 2012 The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**, 49–54. (doi:10.1038/nature11413)
28. Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, Brenner S, Ragsdale CW, Rokhsar DS. 2015 The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* **524**, 220–224. (doi:10.1038/nature14668)
29. Tadepally HD, Burger G, Aubry M. 2008 Evolution of C2H2-zinc finger genes and subfamilies in mammals: species-specific duplication and loss of clusters, genes and effector domains. *BMC Evol. Biol*. **8**, 176. (doi:10.1186/1471-2148-8-176)
30. Sluder AE, Mathews SW, Hough D, Yin VP, Maina CV. 1999 The nuclear receptor superfamily has undergone extensive proliferation and diversification in Nematodes. *Genome Res.* **9**, 103–120. (doi:10.1101/gr.9.2.103)
31. Sharman AC. 1999 Some new terms for duplicated genes. *Semin. Cell Dev. Biol*. **10**, 561–563. (doi:10.1006/scdb.1999.0338)
32. Holland P. 1999 Gene duplication: past, present and future. *Semin. Cell. Biol*. **10**, 541–547. (doi:10.1006/scdb.1999.0335)
33. Faas L, Isaacs HV. 2009 Overlapping functions of Cdx1, Cdx2, and Cdx4 in the development of the amphibian *Xenopus tropicalis*. *Dev. Dyn.* **238**, 835–852. (doi:10.1002/dvdy.21901)
34. Marlétaz F, Maeso I, Faas L, Isaacs HV, Holland PWH. 2015 Cdx ParaHox genes acquired distinct developmental roles after gene duplication in vertebrate evolution. *BMC Biol*. **13**, 56. (doi:10.1186/s12915-015-0165-x)
35. Williams NA, Holland PW. 2000 An amphioxus Emx homeobox gene reveals duplication during vertebrate evolution. *Mol. Biol. Evol.* **17**, 1520–1528. (doi:10.1093/oxfordjournals.molbev.a026251)
36. Peel AD, Telford MJ, Akam M. 2006 The evolution of hexapod engrailed-family genes: evidence for conservation and concerted evolution. *Proc. R. Soc. B* **273**, 1733–1742. (doi:10.1098/rspb.2006.3497)
37. Steinke D, Salzburger W, Braasch I, Meyer A. 2006 Many genes in fish have species-specific asymmetric rates of molecular evolution. *BMC Genomics* **7**, 20. (doi:10.1186/1471-2164-7-20)
38. Fares MA, Byrne KP, Wolfe KH. 2006 Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of Saccharomyces species. *Mol. Biol. Evol*. **23**, 245–253. (doi:10.1093/molbev/msj027)
39. Felsenstein J. 1978 Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Biol*. **27**, 401–410. (doi:10.1093/sysbio/27.4.401)
40. Simeone A, Acampora D, Gulisano M, Stornaiuolo A, Boncinelli E. 1992 Nested expression domains of four homeobox genes in developing rostral brain. *Nature* **358**, 687–690. (doi:10.1038/358687a0)
41. Holland P, Ingham P, Krauss S. 1992 Mice and flies head to head. *Nature* **358**, 627–628. (doi:10.1038/358627a0)
42. Plouhinec J-L *et al.* 2003 The mammalian Crx genes are highly divergent representatives of the Otx5 gene family, a gnathostome orthology class of orthodenticle-related homeogenes involved in the differentiation of retinal photoreceptors and circadian entrainment. *Mol. Biol. Evol.* **20**, 513–521. (doi:10.1093/molbev/msg085)
43. Quiring R, Walldorf U, Kloter U, Gehring WJ. 1994 Homology of the eyeless gene of Drosophila to the small eye gene in mice and Aniridia in humans. *Science* **265**, 785–789. (doi:10.1126/science.7914031)
44. Manousaki T, Feiner N, Begemann G, Meyer A, Kuraku S. 2011 Co-orthology of Pax4 and Pax6 to the fly eyeless gene: molecular phylogenetic, comparative genomic, and embryological analyses. *Evol. Dev.* **13**, 448–459. (doi:10.1111/j.1525-142X.2011.00502.x)
45. Matsushita T, Yamaoka T, Otsuka S, Moritani M, Matsumoto T, Itakura M. 1998 Molecular cloning of mouse paired-box-containing gene (Pax)-4 from an islet β cell line and deduced sequence of human Pax-4. *Biochem. Biophys. Res. Commun.* **242**, 176–180. (doi:10.1006/bbrc.1997.7935)
46. May RM. 1986 Biological diversity: how many species are there? *Nature* **324**, 514–515. (doi:10.1038/324514a0)
47. Mutanen M, Wahlberg N, Kaila L. 2010 Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. *Proc. R. Soc. B* **277**, 2839–2848. (doi:10.1098/rspb.2010.0392)
48. Regier JC *et al.* 2015 A molecular phylogeny for the oldest (nonditrysian) lineages of extant Lepidoptera, with implications for classification, comparative morphology and life-history evolution. *Syst. Entomol.* **40**, 671–704. (doi:10.1111/syen.12129)
49. Després L, David J-P, Gallet C. 2007 The evolutionary ecology of insect resistance to plant chemicals. *Trends Ecol. Evol.* **22**, 298–307. (doi:10.1016/j.tree.2007.02.010)
50. Fürstenberg-Hägg J, Zagrobelny M, Jørgensen K, Vogel H, Møller BL, Bak S. 2014 Chemical defense balanced by sequestration and de novo biosynthesis in a Lepidopteran specialist. *PLoS ONE* **9**, e108745. (doi:10.1371/journal.pone.0108745)
51. Chai CL, Zhang Z, Huang FF, Wang XY, Yu QY. 2008 A genomewide survey of homeobox genes and identification of novel structure of the Hox cluster in the silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol*. **38**, 1111–1120. (doi:10.1016/j.ibmb.2008.06.008)
52. Ferguson L, Marlétaz F, Carter J-M, Taylor WR, Gibbs M, Breuker CJ, Holland PWH. 2014 Ancient expansion of the hox cluster in Lepidoptera generated four homeobox genes implicated in extra-embryonic tissue formation. *PLoS Genet*. **10**, e1004698. (doi:10.1371/journal.pgen.1004698)

53. The Heliconius Genome Consortium *et al.* 2012 Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98. (doi:10.1038/nature11041)
54. Marlétaz F, Paps J, Maeso I, Holland PWH. 2014 Discovery and classification of homeobox genes in animal genomes. *Methods Mol. Biol*. **1196**, 3–18. (doi:10.1007/978-1-4939-1242-1_1)
55. Paps J, Xu F, Zhang G, Holland PWH. 2015 Reinforcing the egg-timer: recruitment of novel lophotrochozoa homeobox genes to early and late development in the pacific oyster. *Genome Biol. Evol.* **7**, 677–688. (doi:10.1093/gbe/evv018)
56. Duboule D. 1994 Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev. Suppl*. **1994**, 135–142.
57. Booth HAF, Holland PWH. 2007 Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line. *Gene* **387**, 7–14. (doi:10.1016/j.gene.2006.07.034)
58. Booth HAF, Holland PWH. 2004 Eleven daughters of NANOG. *Genomics* **84**, 229–238. (doi:10.1016/j.ygeno.2004.02.014)
59. Holland PWH, Booth HAF, Bruford EA. 2007 Classification and nomenclature of all human homeobox genes. *BMC Biol*. **5**, 47. (doi:10.1186/1741-7007-5-47)
60. Zhong Y-F, Holland PW. 2011 The dynamics of vertebrate homeobox gene evolution: gain and loss of genes in mouse and human lineages. *BMC Evol. Biol.* **11**, 169–237. (doi:10.1186/1471-2148-11-169)
61. Töhönen V *et al.* 2015 Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. *Nat. Commun.* **6**, 8207. (doi:10.1038/ncomms9207)
62. Leidenroth A, Hewitt JE. 2010 A family history of DUX4: phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene. *BMC Evol. Biol*. **10**, 364. (doi:10.1186/1471-2148-10-364)
63. Maeso I, Dunwell TL, Wyatt CDR, Marlétaz F, Vető B, Bernal JA, Quah S, Irimia M, Holland PWH. 2016 Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals. *BMC Biol.* **14**, 47. (doi:10.1186/s12915-016-0267-0)
64. Yan L *et al.* 2013 Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol*. **20**, 1131–1139. (doi:10.1038/nsmb.2660)
65. Takeda J, Seino S, Bell GI. 1992 Human Oct3 gene family: cDNA sequences, alternative splicing, gene organization, chromosomal location, and expression at low levels in adult tissues. *Nucleic Acids Res* **20**, 4613–4620. (doi:10.1093/nar/20.17.4613)
66. Liedtke S, Enczmann J, Waclawczyk S, Wernet P, Kögler G. 2007 Oct4 and its pseudogenes confuse stem cell research. *Cell Stem Cell* **1**, 364–366. (doi:10.1016/j.stem.2007.09.003)
67. Xue Z *et al.* 2014 Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**, 593–597. (doi:10.1038/nature12364)
68. Madissoon E *et al.* 2016 Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. *Sci. Rep*. **6**, 28995. (doi:10.1038/srep28995)
69. Tsai IJ *et al.* 2013 The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* **496**, 57–63. (doi:10.1038/nature12031)